A Misdirected Principle with a Catch: Explicability for AI

Robbins, Scott

DOI
10.1007/s11023-019-09509-3

Publication date
2019

Document Version
Final published version

Published in
Minds and Machines: journal for artificial intelligence, philosophy and cognitive sciences

Citation (APA)
Robbins, S. (2019). A Misdirected Principle with a Catch: Explicability for AI. *Minds and Machines: journal for artificial intelligence, philosophy and cognitive sciences*, *29*(4), 495-514. https://doi.org/10.1007/s11023-019-09509-3

Minds and Machines
https://doi.org/10.1007/s11023-019-09509-3

A Misdirected Principle with a Catch: Explicability for AI

Scott Robbins[1]

Received: 6 June 2019 / Accepted: 5 October 2019

Abstract
There is widespread agreement that there should be a principle requiring that artificial intelligence (AI) be 'explicable'. Microsoft, Google, the World Economic Forum, the draft AI ethics guidelines for the EU commission, etc. all include a principle for AI that falls under the umbrella of 'explicability'. Roughly, the principle states that "for AI to promote and not constrain human autonomy, our 'decision about who should decide' must be informed by knowledge of how AI would act instead of us" (Floridi et al. in Minds Mach 28(4):689–707, 2018). There is a strong intuition that if an algorithm decides, for example, whether to give someone a loan, then that algorithm should be explicable. I argue here, however, that such a principle is misdirected. The property of requiring explicability should attach to a particular action or decision rather than the entity making that decision. It is the context and the potential harm resulting from decisions that drive the moral need for explicability—not the process by which decisions are reached. Related to this is the fact that AI is used for many low-risk purposes for which it would be unnecessary to require that it be explicable. A principle requiring explicability would prevent us from reaping the benefits of AI used in these situations. Finally, the explanations given by explicable AI are only fruitful if we already know which considerations are acceptable for the decision at hand. If we already have these considerations, then there is no need to use contemporary AI algorithms because standard automation would be available. In other words, a principle of explicability for AI makes the use of AI redundant.

Keywords Ethics of AI · Explicability · Explainable AI · Meaningful human control · Artificial intelligence

✉ Scott Robbins
scott@scottrobbins.org

[1] Delft University of Technology, Jaffalaan 5, 2628 BX Delft, The Netherlands

Published online: 15 October 2019

1 Introduction

It is rare to see large numbers of ethicists, practitioners, journalists, and policy-makers agree on something that should guide the development of a technology. Yet, with the principle requiring that artificial intelligence (AI) be explicable, we have exactly that. Microsoft, Google, the World Economic Forum, the draft AI ethics guidelines for the EU commission, etc. all include a principle for AI that falls under the umbrella of 'explicability'. The exact wording varies. Some talk of 'transparency', others of 'explainability', and still others of 'understandability'. Finally, Floridi et al. call for a principle of 'explicability' for AI which claims that when systems are powered by AI, humans should be able to obtain "a factual, direct, and clear explanation of the decision-making process" (Floridi et al. 2018).

The intuition that an algorithm should be capable of explaining itself is strong—especially algorithms operating in morally significant contexts. Frank Pasquale's Black Box Society (2015) provides examples of decisions made about us by algorithms for which we are not offered an explanation. It is unfair that we can receive a low credit score, end up on a police watch list, get higher prison sentences, etc. without explanation about the considerations that led to those decisions. If algorithms are used to make decisions in these contexts, there should be explanations about how they arrived at a specific decision.[1] Floridi et al. argue that AI will constrain rather than promote human autonomy unless we have the "knowledge of how AI would act instead of us" (2018, p. 700).

Getting algorithms to provide us with explanations about how a particular decision was made allows us to keep 'meaningful human control' over the decision. That is, knowing why a particular decision was reached by an algorithm allows us to accept, disregard, challenge, or overrule that decision.[2] 'Meaningful human control' was originally used as a principle for lethal autonomous weapons systems: "humans not computers and their algorithms should ultimately remain in control of, and thus morally responsible for relevant decisions about (lethal) military operations" (Article 36 2015). 'Meaningful human control' is now being used to describe an ideal that all AI should achieve if it is going to operate in morally sensitive contexts (see e.g. Robbins 2019; Santoni de Sio and van den Hoven 2018). A principle of explicability, then, is a *moral* principle that should help bring us closer to acceptable uses of algorithms. The question then is: does a principle of explicability overcome ethical issues associated with the use of algorithms?

In what follows, I will argue that principles requiring that AI be explicable are misguided. Not only would such a requirement trade off the power of AI in terms of performance, but such a requirement assumes that we have a list of considerations that are acceptable for a given decision. I argue that such a list would preclude the

[1] Robbins and Henschke make the important point that this argument can be turned on its head: "The solution, therefore, is to use such algorithms for specific situations in which it is acceptable to not have an explanation" (Robbins and Henschke 2017).

[2] This is not the only conception of meaningful human control in the literature. More will be said about this in what follows.

use of machine learning algorithms. Of more philosophical importance is that the property of 'requiring explicability' is incorrectly applied to AI. The real object in need of the property of 'requiring explicability' is the result of the process—not the process itself. We do not require everyone capable of making a decision to be able to explain every decision they make. Rather, we require them to provide explanations when the decisions they have made require explanations. For AI we should take a similar approach.

Instead of trying to have our cake and eat it too (having powerful AI that can explain its decisions), we should be deciding which decisions require explanations. Knowing that a specific decision requires an explanation (e.g. declining a loan application) gives us good reason *not* to use opaque AI (e.g. machine learning) for that decision. Any decision requiring an explanation should not be made by machine learning (ML) algorithms. Automation is still an option; however, this should be restricted to the old-fashioned kind of automation whereby the considerations are hard-coded into the algorithm. Luckily for the ML community, there are many decisions that benefit society without requiring explanations.

2 Calls for a Principle of Explicability for AI

It would be shadowboxing to argue that a principle of explicability for AI is unnecessary if there were no proposals for such a principle. In this section, I highlight some examples of the many calls for such a principle by academics, NGOs, corporations, etc. It should be clear that explicability is considered to be an important part of achieving so-called 'ethical', 'responsible', 'trustworthy', etc. AI.

Before highlighting the many examples of calls for a principle of explicability for AI, it is important to distinguish between the usefulness of explicable AI and a requirement that AI be explicable. I do not argue against the idea that explicable AI could be useful in certain contexts; rather, I will argue against a principle requiring that AI be explicable. For example, if someone were to have an ML algorithm that was highly accurate with regard to making predictions about the weather, there may be some desire to have that algorithm explain itself. This desire would not be based on the idea that it is wrong to use the decisions made by the ML without explanation; rather, knowing what considerations were used by the ML for its decision may increase our knowledge about the weather. This example is in contrast with the examples used by those proposing a principle of explicability for AI. ML used for medical diagnosis (de Bruijne 2016; Dhar and Ranganathan 2015; Erickson et al. 2017), judicial sentencing (Berk et al. 2016; Barry-Jester et al. 2015), and predictive policing (Ahmed 2018; Ensign et al. 2017; Joh 2017) are just a few of many real-world examples. Using the decisions of ML algorithms in these contexts without explanation is wrong, so the argument goes, unless that ML algorithm is explicable.

One reason that using inexplicable decisions in morally sensitive contexts like the ones listed above is wrong is that we must ensure that the decisions are not based

on inappropriate[3] considerations. If a predictive policing algorithm labels people as criminals and uses their skin color as an important consideration then we should not be using that algorithm. If the algorithm is not explicable, then this consideration may be used without our knowledge. The opacity of the algorithm prevents us from knowing whether it is unethically biased.

One of the main reasons that AI, and ML specifically, is the target in calls for a principle of explicability is that these algorithms are opaque. The inputs used for ML algorithms[4] are translated into a machine-readable format (1 s and 0 s) and then based on the patterns those 1 s and 0 s have a path is taken through a series of hidden layers. The data used to train this algorithm will have given each of the many paths that an input could take a probability corresponding to the resulting classification. Although many researchers are working to make this process explicable, little progress has been made (see e.g. Gilpin et al. 2018; Kuang 2017; Wachter et al. 2017). Those who have had some success can only give us educated guesses based on many results. In a nutshell, they are using algorithms to analyze the results for patterns that may tell us something about the reasons used by the target ML algorithm.

In short, we do not know the reasons for a specific ML algorithm decision. Combine this fact with using ML algorithms for decisions that the as having moral significance (i.e. decisions which could result in harm-broadly construed to include rights violations) and we have an ethically problematic situation. An algorithm used, for example, to accept or reject your loan request will significantly affect you. A rejection could cause you and your partner significant distress and change the course of your life. It is exactly this type of situation that motivated the European Union to include in the General Data Protection Regulation (GDPR) what many have interpreted as a 'right to explanation' when fully automated decisions significantly affect someone:

> *the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her*[5]

It is intuitive that, when an ML algorithm makes a decision about us that has a morally 'significant' effect, it should be able to 'explain' itself. This intuition has led many to propose that a principle of AI is that it should be explicable. Below is a sample of the academics, non-governmental organizations, and large technology companies who have an AI principle that can be interpreted to be an explicability principle.

[3] Inappropriate captures both considerations that are unethical (e.g. race) and clearly irrelevant (e.g. your astrological sign). Both are inappropriate and could lead to unethical outcomes.

[4] I specifically discuss deep learning algorithms here. Note that other ML algorithms using different methods exist (e.g. evolutionary algorithms).

[5] GDPR Recital 71. The full text can be found at https://gdpr-info.eu/recitals/no-71/. Some have argued that no such right can be derived (Wachter et al. 2016).

Luciano Floridi, for example, outlined a framework for a 'Good AI Society'. In that framework he and his colleagues explicitly call for AI systems that make 'socially significant decisions' to be explicable:

> *Develop a framework to enhance the explicability of AI systems that make socially significant decisions. Central to this framework is the ability for individuals to obtain a factual, direct, and clear explanation of the decision-making process, especially in the event of unwanted consequences.* (Floridi et al. 2018, p. 702)

NGOs including the Public Voice (established by the Electronic Privacy Information Center) and the Future of Life Institute have also called for principles of explainability for AI.[6] The Public Voice, in their list of AI Universal Guidelines, has a right to transparency which states:

> *All individuals have the right to know the basis of an AI decision that concerns them. This includes access to the factors, the logic, and techniques that produced the outcome.* (AI Universal Guidelines—thepublicvoice.org 2018)

And the Future of Life Institute includes two transparency principles in their AI Principles:

> *Failure transparency: If an AI system causes harm, it should be possible to ascertain why.*
> *Judicial transparency: Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.* (AI Principles 2017)[7]

Microsoft's current CEO Satya Nadella called for a transparency requirement in an op-ed to the online magazine Slate:

> *A.I. must be transparent: We should be aware of how the technology works and what its rules are. We want not just intelligent machines but intelligible machines. Not artificial intelligence but symbiotic intelligence. The tech will know things about humans, but the humans must know about the machines. People should have an understanding of how the technology sees and analyzes the world. Ethics and design and in hand.* (Nadella 2016)

And Google claims that they will "design AI systems that provide appropriate opportunities for feedback, relevant explanations, and appeal" (AI at Google 2018).

Last but not least, James Bridle in his book *The New Dark Age: Technology and the end of the Future* calls for a fourth principle of robotics (to add to Asimov's first

[6] For other examples of principles which could be interpreted as requiring AI to be explainable see UNI Global Union (2018), the Partnership on AI (2019). There are sure to be more.

[7] It is unclear why the judicial context gets special attention here. While the judicial context is an especially morally salient one, it is none more so than medical or policing contexts.

three): "a robot—or any other intelligent machine—must be able to explain itself to humans" (2019).[8]

While it is not possible to claim that this sample of principles, and the many others I did not mention, all amount to the same thing, they do all call for AI to be explicable. To be sure, I do not think it is the intent of the authors of these principles to require *all* AI to be explicable; however, the way that the principles are written this requirement would either apply to all AI or it would be unclear when it would have to be applied or not. In some of the examples above the principles call for transparency of AI. Although transparency and explicability are not synonymous, when transparency is used with respect to the transparency of the reasons for the AI-generated decision, this amounts to explicability. Others have called for transparency principles which are not the same as explicability. Instead, what they mean by transparency is transparency of the sourcing and usage of training data or transparency of other parts of the development and implementation of AI.[9] One can support this kind of transparency without supporting a principle of explicability. One may, for example, be transparent with regard to the training data used for the algorithm without being able to provide an explanation regarding a particular decision made by that algorithm. This kind of transparency would go some way toward ensuring that algorithms will work for a diverse set of people (e.g. ensuring that the training data was not solely made up of the data regarding white males).

The many examples highlighted above are there to make it clear that there are many calls for AI to be explicable. Indeed, not just calls, but demands for a principle that would require AI to be explicable. It is the purpose of this paper to argue that such a principle is misguided.

3 The Why, Who, and What of an Explicability Principle for AI

3.1 What is Explicability For?

Before getting to what explicability is and who it is for, we must understand what the purpose is for a principle of explicability for AI. This will go some way towards understanding what explicability is and who it is for. I argue that a principle of explicability is primarily for the maintaining of meaningful human control over algorithms. The idea is that an explanation of an algorithm's output will allow a human being to have meaningful control over the algorithm—enabling the ascription of moral responsibility to that human being (or set of human beings). With an explanation of the algorithm's decision, it is possible for human beings to accept, disregard, challenge, or overrule that decision. The Center for a New American Security (CNAS), for example, writes that it is necessary that "human operators are

[8] It must be noted that Asimov originally had four total laws-meaning that Bridle's would be a fifth, not a fourth. He added a 'zeroeth law' to precede the others which stated: "a robot may not harm humanity, or, by inaction, allow humanity to come to harm.".

[9] See e.g. Whittaker et al. (2018).

making informed, conscious decisions about the use of weapons" and that "human operators have sufficient information to ensure the lawfulness of the action they are taking…".[10]

There are, however, other features of meaningful human control that would not be captured by explicability. Meaningful human control over autonomous driving systems may not require human beings to have any say over a particular decision because of the psychological limitations of the human driver to gain cognitive awareness in time to act (Heikoop et al. 2019). Santoni de Sio and van den Hoven (2018) argue that meaningful human control occurs when algorithms meet 'track' and 'trace' conditions. We must be able to trace responsibility for the outcomes of algorithms back to human beings. The decisions of algorithms must also track human values. While I use a specific conception of meaningful human control (i.e. giving humans the ability to accept, disregard, challenge, or overrule an AI algorithm's decision), I am not arguing that this conception is the best one. Rather, this is the conception that I argue is implicit when one requires that AI be explicable.

We must keep in mind that an explicability principle for AI is ethical in nature. The starting point for these lists is that there are ethical problems associated with algorithms. If the design and development of algorithms follow a particular set of principles, then, it is believed that the resulting algorithm will be 'good', 'trustworthy', or 'responsible'. So, a principle of explicability is an attempt to overcome some ethical issues unique to algorithms.

Ethical value is to be contrasted to the epistemic value explicable AI might provide. Explicable AI may be extremely valuable to researchers and others who would be able to use explanations to better understand their domain. Garry Kasparov, for example, may find an explanation of a particular chess move made by an algorithm beneficial for his own ability to play chess.[11] A doctor may find an explanation useful to better understand how to diagnosis a particular disease. This epistemic value of explicability for AI is not under dispute. In these cases, we are not harmed by the opacity of the algorithm's decision-making process. A principle of explicability, in contrast, is ethical in that it is about preventing harm (broadly construed) that could occur due to the opacity of the algorithm.

What is the ethical issue that is giving rise to this principle? One candidate is the issue of understanding what went wrong if something harmful happens as a consequence of the algorithm. For example, if a self-driving car swerves into a barrier killing its passenger(s) then it would be helpful to have an explanation of what caused this to happen in order to prevent it happening in the future. While a principle of explainability would help with this, it does not capture the full range of ethical issues that explicability aims to overcome. For example, if someone is incorrectly

[10] For other documents with similar features for meaningful human control see e.g. United States Department of Defense (2012) and Horowitz and Scharre (2015). For a helpful overview of the common themes involved in discussions about meaningful human control see Ekelhof (2019).

[11] A good, recent, example of this is the growing discussion about Move 37 by AlphaGo during its game with Lee Sedol (Metz 2016).

denied a loan by an algorithm how will we know that something harmful has happened so that we can demand an explanation?

This points to the ethical issue of ensuring that the outputs of algorithms are not made based upon ethically problematic or irrelevant considerations. We expect, for example, a rejection for a loan not to be based on the color of the applicant's skin (or a proxy thereof). An explanation of the algorithm's decision can allow for someone to accept, disregard, challenge, or overrule the rejection. This gives meaningful control of the decision to human beings. This goes above and beyond the stipulation that some particular human is responsible for the algorithm's decisions. This provides a human with the information they need in order to exercise that control.

Explicability, therefore, is an attempt to maintain meaningful human control over algorithms. Only human beings can be held morally accountable so it should be human beings that are in control over these decisions (see e.g. Johnson 2006). If a human being has an explanation of the algorithm's decision, then it is possible for that human being to accept, disregard, challenge, or overrule that decision.

3.2 Who is Explicability for?

How the requirement that AI be explicable is understood depends upon who will receive the explanation. A medical diagnosis algorithm that classifies someone as having a brain tumor might, for example, provide a heat map of which parts of the brain scan most contributed to the diagnosis. This 'explanation' would probably be useless to a patient—or to anyone else without very specific medical training. However, if the goal is that the algorithm is under 'meaningful human control' then we are not concerned with the patient's understanding of the explanation.

Just as with any diagnosis, we trust that our physician is making a justified decision in line with current medical practice. The physician should be ultimately responsible for the brain tumor diagnosis and therefore it is the physician who should be able to evaluate the explanation. Remember that the purpose of the explanation is to overcome an ethical problem; namely, to establish meaningful human control over that decision by allowing one to confirm that the reasons for a decision are in line with domain-specific norms and best practices.

To illustrate, let us say that an algorithm rejects a loan application. This algorithm is able to provide an explanation in the form of considerations that played a factor in its rejection. One of those considerations was the fact that the application included a high debt-to-income ratio. To the applicant, this is interesting to know but it would be quite unclear whether their debt-to-income ratio was at a level that justified its factoring in on a decision to reject their loan application. Only those with relevant domain-specific knowledge would be able to evaluate whether this particular debt-to-income ratio should factor into a decision to reject the loan. This only gets more complicated as more considerations factor into algorithmic decisions.

To achieve the ethical goal of a principle of explicability the explanation provided by an algorithm should enable a human being to have meaningful control over the decisions the algorithm makes. This means that the person using the algorithm is the person that the explanation should be directed towards—not the person subject

to the decision of the algorithm (although those two roles may be filled by the same person). While the person subject to the algorithm's outputs may be interested to know the explanation (and in some cases should be provided with it in order to achieve other ethical goals),[12] this does not establish meaningful human control over the algorithm's output.

3.3 Artificial Intelligence

'Artificial Intelligence' is an overused phrase that signifies many things. Explanation also has many uses depending on the context. We have had artificially intelligent systems for decades that did not result in any calls for explanation. This is mainly because what is known as good old-fashioned AI (GOFAI) is simply a set of explicitly coded rules in the form of a decision tree that allows for the automation of processes. For example, if you wanted to automate the decision on which move to make in chess it may look like this:

```
If (first move of game) then move random pawn 2 spaces forward
Else if (king is in check) then (move king to non-checked space)
Else if (possible to achieve checkmate) then (achieve checkmate)
Else if (possible to put king in check) then (put king in check)
Else if (possible to take an opposing players piece) then (take piece)
Else (move piece at random)
```

This is clearly a terrible algorithm for deciding your next chess move—a much more sophisticated algorithm designed using GOFAI could be achieved. However, this kind of automation is inherently explicable because the code makes the reasons for a resulting decision explicit. Opacity with regard to this type of automation would only occur if the institutions doing the automating did not want people to know how the decisions are being made (see e.g. Pasquale 2015).

GOFAI is in contrast to AI that falls under the umbrella of machine learning (ML). The GOFAI approach is limited by what the designers of the algorithm could think of. Novel situations may result in terrible decisions by the AI. ML is one approach to overcome such limitations. In a nutshell, ML attempts to use statistical methods to allow an algorithm to 'learn' every time it 'tries' to achieve its specified goal. Each attempt, whether it fails or succeed, will result in the algorithm updating its statistical probabilities that correlate to features of the input.[13]

An ML algorithm could be trained to play chess by playing many times without explicit rules given by humans. The ML algorithm may play at random the first

[12] Most notably the goal of *actionable recourse*: the ability to contest incorrect decisions or to understand what could be changed in order for the data subject to achieve a more desirable result (Wachter et al. 2017; Ustun et al. 2019).

[13] For a nice overview of machine learning methods and trends see Jordan and Mitchell (2015).

time—losing very easily. At the end of the game, we would tell the AI that it lost. The next game the AI would play slightly differently. Over hundreds, thousands, or even millions of games the AI would be very well trained to play the game of chess. The resulting trained ML algorithm would be opaque with regard to its reasoning for any given move.

Is it acceptable that the algorithm makes decisions that are not explicable? If you share my intuition that there is no problem here, it may stem from the fact that the outcomes of these 'chess move' decisions cannot result in harm. A terrible chess move may result in the loss of the chess game, but life, limb, reputation, and property are not at stake. An AI making decisions in other contexts, such as medical diagnosis and judicial sentencing, could cause real harm.

The point here is to show that the principle of explicability is important due to the rise of algorithms using ML or other methods that are opaque with regard to how the algorithm reaches a particular decision. If we are simply using automated processes (e.g. GOFAI) then explicability is only a problem if the developer intentionally obfuscates the explanation. In these cases an explanation is readily available to developers and companies; however, they do not see it in their interest to reveal that explanation to the public. While not addressed here, this problem is very important (see Pasquale 2015).

3.4 Explicability

So if one is using an ML algorithm for decisions that could result in harm and responsibly wants to adhere to a set of principles that includes a principle of explicability, what is one to do? First, one would need to know what is being demanded by a principle of explicability. That is, what is an explanation that would satisfy the principle?

First, we could be demanding a causal explanation for a particular outcome/action/decision. For example, when Google's image classification algorithm classified two young black people as gorillas there was an outcry and much embarrassment for Google (Kasperkevic 2015). If Google were to explain the algorithm's classification by saying that "features of the image input correlated highly with training images classified as gorillas" I doubt that anyone would be satisfied. We are not concerned with how the algorithm classifies images in general. Rather, we want to know why the label 'gorillas' was applied to a specific image by the algorithm. In other words, we demand to know the specific features of the image that contributed to the labeling.

Scientific explanations also give us answers to *how* things happened. However, we do not want to know the *how*; rather, we want to know the *why*. I do not want to know how my daughter hit her brother: "I raised my right arm and moved it forward at high velocity", but the why: "he took my favorite stuffed animal from me." The latter *why* explanation is an explanation that provides the reason(s) that a particular action was taken. This reason or reasons may or may not morally justify the action. These reasons are precisely what we want to evaluate. In the case of ML we could

get an explanation like the following excerpt used to describe how DeepMind's AlphaGo chooses its next move:

> *At the end of the simulation, the action values and visit counts of all traversed edges are updated. Each edge accumulates the visit count and mean evaluation of all simulations passing through that edge is the leaf node from the ith simulation, and I(s, a, i) indicates whether an edge (s, a) was traversed during the ith simulation. Once the search is complete, the algorithm chooses the most visited move from the root position* (Silver et al. 2016)

This, if you are a person with the requisite knowledge to understand it, is an explanation of the *how* for a particular move in the game of Go made by the algorithm-driven process. It says nothing about the particular features of that move which contributed to the decision to make the move. One could attempt to provide a justification for a particular move made by the algorithm by referencing the effectiveness of the algorithm itself: "the move chosen by the algorithm is a good move because the algorithm has proven to be very good at the game of Go". We can see that this is an unsatisfying explanation when we apply it to a different context. If the best heart surgeon in the world were to leave a sponge in the patient and a nurse were to ask: "why did she leave the sponge in the patient?" and someone were to respond "it was good to leave the sponge there because the decision was made by the best surgeon in the world." What we really want with an explanation are all (and only) the considerations important for their contribution to a particular decision—considerations that a human could use to determine whether a particular algorithmic decision was justified.

We could give a general explanation of sorts for opaque algorithms in any context. Why did the ML algorithm decide to label a convicted criminal as high-risk? Because data used as an input to the algorithm correlated with features of data used to train the algorithm that were tagged as having a high risk. While this is an explanation, it clearly falls short of what is desired by the principles highlighted above. What is really desired is an explanation that would provide a human with information that could be used to determine whether the result of the algorithm was *justified*.

An explanation may justify a particular decision or it may not, and, a decision may be justified by reasons that do not feature in an explanation of that decision (see e.g. Dancy 2004, ch. 5; Darwall 2003). If, for example, I were to make a move in chess because I thought that it would make the board more balanced (in terms of aesthetics) we would have an explanation for the move that I made that failed to justify the move. However, that move may also have been the best move I could have made—making the move justified. While it was a great chess move, I doubt anyone would take my advice on a future move—nor should we if we knew that an algorithm was using board balance as a consideration in favor of a particular move. This shows that we cannot simply look to the decision itself and ask whether that decision was justified or not. An algorithm may flag someone as a dangerous criminal who in fact happens to be a dangerous criminal—justifying the algorithm's classification. However, if the consideration leading to that classification was the person's race then we have an explanation that fails to justify the decision whether the decision was correct.

In short, what is desired is an explanation providing the considerations that contributed to the result in question. This gives a human being the information needed in order to accept, disregard, challenge, or overrule the decision. In the same way that a police officer might claim in court that a particular criminal is high-risk, and the judge asks for the considerations used to justify such a label, we want the algorithm to justify itself in reference to the considerations used.

A justification for this 'high risk' label given by the police officer might be that while in custody the criminal threatened to do much more harm once she was free. The judge may accept this as a good justification and sentence the criminal to the maximum allowable prison sentence. If, on the other hand, the police officer justified this label by saying that the criminal was really dark-skinned and menacing looking, then the judge (hopefully) would reject the police officer's label of 'high-risk'. If an algorithm was delegated the task of labeling criminals as 'high-risk' and did so as a result of race, then we would want the judge to know that so that she could reject the algorithm's decision. A technical, causal, or scientific explanation does not allow the judge to have meaningful human control over the algorithm.

4 Current Approaches to Explicable AI

Having a principle requiring that AI be able to explicable means that there must be methods for which an algorithm can give an explanation for its decision. Here I do not focus on intrinsically explainable algorithms (like the GOFAI approach above). Instead, I focus on the ML algorithms that are the reason for introducing a principle of explicability for AI in the first place.

There has been much work in achieving explainable AI. They can be classified into two broad approaches. The first is offers 'model-centric explanations' and the second offers 'subject-centric explanations' (Edwards and Veale 2017, p. 22). Model-centric explanations aim to provide the information that is known about the algorithm in order to better understand the algorithm—enabling users to better understand how to use the algorithm. The information that provided relates to the data the algorithm was trained on, how the algorithm was tested for bias, the intentions of the designers, performance metrics, etc. The idea is that knowing all of this other information about the algorithm may allow society to "make informed choices regarding usage, implementation, and regulation of these machines" (Robbins 2019).

While this approach to explainable AI is interesting, it does not really capture what is meant by 'explicability'. We do not have the considerations that played a factor in the resulting decision. At best we have guestimates or maybe a justified belief that the algorithm will work in the given context because it has performed well in similar contexts and the input is relevantly similar to data used during the training phase of the algorithm. This does not overcome the ethical problem resulting in important decisions made by algorithms. However, it may significantly help society decide on the acceptability of using a specific algorithm for a specific purpose.

The 'subject-centric' explanations are an attempt to zoom in on the input (the subject) and understand what it is about it that caused the specified decision. For

example, an explanation of a loan rejection may be that the person who requested the loan has a debt-to-income ratio that is always classified as a rejection by the algorithm. While there may be other considerations that would also contribute to a rejection, the debt-to-income ratio could be seen as a sufficient condition for rejection. In this clear cut case, the explanation would help humans decide whether the decision was justified—and therefore satisfy the type of explanation discussed in the previous section. Unfortunately ML decisions are rarely going to be this simple. The more data fed into the algorithm as input makes the output that much harder to explain. Many variables may need to be modified in order to change the resulting classification—making it increasingly unlikely that a satisfactory explanation is provided.

5 Three Misgivings about Explicable AI

There are three major misgivings I have regarding the principle of explicability for AI. The first is with regard to where the property of 'requiring explicability' is placed. I argue that we do not normally place such a property on the process which results in a decision; rather, we place that property onto the decision itself. Second, there seem to be many implementations of AI in situations of low to no risk (in terms of harm). It is unreasonable that the decisions resulting from AI in these situations should be required to provide explanations. Finally, in situations of high risk there is a catch-22 for those who wish to use ML: If ML is being used for a decision requiring an explanation then it must be explicable AI and a human must be able to check that the considerations used are acceptable, but if we already know which considerations should be used for a decision, then we don't need ML.

5.1 Explicability of the Decision versus Decision-Maker

The mistake with requiring that AI be explicable is that it places the requirement of explicability onto the decision-maker rather than the decision itself. Some calls for a principle of explicability allude to this when they add the qualifier resembling 'when the decision made by the AI significantly affects a person'. This is an acknowledgment that the property of 'requiring an explanation' really applies to the decision itself—not the entity making that decision.

When my daughter hits her brother, I would reasonably demand that she explain her decision to act in that way. She has significantly impacted her brother because she has directly caused him pain. In contrast, when my daughter suddenly starts to dance and I ask her why, she would (and has done many times) shrug her shoulders and say, "I don't know". I, of course, am not mad at her for her lack of explanation. The reason is that one action requires an explanation, and the other does not. The first action resulted in harm, thereby 'significantly affecting' a person. The second action is benign. No one is harmed by my daughter's spontaneous dancing. It

would therefore be unreasonable if I were to tell my daughter that everything she did requires a morally justifying explanation[14] or that all children should be 'explicable'.

In short, adding the property 'requiring explicability' to children would be a mistake. It is the action or decision which can/should have the property of requiring explicability. Decisions capable of causing harm (broadly construed) are decisions which require this property. Anyone unable to give an explanation for such a decision is doing wrong.

When discussions about AI and explanation come up, there are some common examples given. Algorithms making decisions about loan applications, criminal sentencing, policing, medical diagnoses, weapons targeting, etc. all get mentioned when discussing the need for algorithms to be able to explain themselves. However, the common element in all of these contexts is that the decisions made in these contexts require explanations that justify those decisions. Whatever the process used to make these decisions there must be an explanation for any given decision.

This is important because using explicability as a principle for AI could force those designing algorithms for decisions or roles that do not require explanation to use less powerful AI like GOFAI. This would significantly constrain many of the great uses of ML algorithms that are not able to explain themselves. For example, ML is often used for credit card fraud detection (see e.g. Morrell 2018). When the algorithm classifies a transaction as fraudulent, this causes the bank to lock the credit card until the customer can confirm that they indeed made the transaction. False positives can, to be sure, be annoying; however, the only thing we really care about is whether the algorithm performs well compared to other methods. Because the role of the algorithm is simply to flag a transaction as fraudulent, the ultimate decision-maker will be the customer herself. I can see no good reason why the ML algorithm should be forced to provide an explanation here.

This is why many of the principles highlighted above include a qualification; namely, that AI must be explicable if the decision will significantly affect someone. This, of course, needs to be specified very clearly in order to separate out the decisions that will trigger the principle and those that do not. Of course, once we do this with any level of specification we are simply deciding what roles, tasks, and decisions require explanations and which ones do not. The principle will no longer have anything to do with artificial intelligence.

5.2 Inexplicable AI for Low-Risk Purposes

In May 2015 Google's AlphaGo algorithm defeated the world champion Go player Ke Jie (France-Presse 2017). The AlphaGo algorithm provided Ke Jie no explanation for any of the moves it made. However, an algorithm deciding which moves to make in the game of Go does not seem problematic because the possible consequences stemming from these decisions are at no risk of causing harm. Many AI

[14] This does not preclude my interest in an explanation in terms of, for example, her desires and preferences. If she simply told me that she "loved dancing" when asked "why" then this may provide me with a reason for entering her into dance class.

and ML applications fall into this category. This is more often than not a result of the algorithm's implementation within a larger process. For example, Cortis is an algorithm that detects voice patterns associated with cardiac arrest (Vincent 2018). The algorithm exists explicitly for the purposes of aiding emergency call operators. The algorithm takes as its input live sound from the calling line. Its output is true if the voice pattern is associated with cardiac arrest and false if it is not. The context of emergency calls is high risk. The operator has legal, as well as moral, responsibility and can make decisions that will save (or end) lives. The addition of the algorithm in this context aids the operator with one specific problem: someone on the other end of the line may be having a heart attack.

This algorithm cannot, however, cause harm. The worst-case scenario is that the algorithm does not identify someone as having a cardiac arrest who is indeed experiencing cardiac arrest. This is regrettable; however, the algorithm not being there would not have changed this outcome.[15] It is an example of an algorithm that should be judged on its accuracy—not its reasons. A principle of explicability would mean this algorithm would not be allowed to operate. This would be unfortunate as it has been shown to detect heart attacks on average 30 s faster than human operators with an accuracy of 93% (human operators have a 73% accuracy rate).

It should be noted that establishing that a particular algorithm has no risk of causing harm would be incredibly difficult to establish. It will often be the case that it is unknown what the possible consequences of algorithmic decisions will be. There would have to be some standard of risk for automated decisions before we allow anyone to claim that their algorithm's decisions cannot cause harm. The point here is to show that there are definitely cases where algorithm's decisions have a low risk of causing harm and a lack of explanation should not preclude its use.[16]

5.3 Catch 22 of Requiring Explicability for AI

In Joseph Heller's Catch 22 Doc Daneeka explains to Yossarian the catch regarding the policy allowing insane people to cease flying bombing missions: "Catch 22. Anyone who wants to get out of combat duty isn't really crazy" (Heller 2011, p. 52). So in order to get out of combat duty, one would have to be insane and to tell their superior that they wished to cease combat duty. Unfortunately, only a sane person would make such a request. There is a similar catch to explainable AI. If ML is being used for a decision requiring an explanation, then it must be explicable AI and a human must be able to check that the considerations used are acceptable. But if we already know which considerations should be used for a decision, then we do not need ML.

An example may help to illustrate: say there is a ML algorithm that is developed to decide whether someone should get a loan (there have been real-world examples

[15] There is a concern that operators may come to think that there is no heart attack unless the algorithm identifies one—resulting in situations where were there not an algorithm they would have identified a heart attack on their own (thanks to Seumas Miller for this concern).

[16] Thanks to an anonymous reviewer for making this point.

of this). This algorithm is opaque and there are justifiably calls for explainable AI in this context. So we pour millions in funding to come up with explainable AI that somehow is just as powerful as the original algorithm.[17] Now when a loan application is processed, there is an explanation spit out by the algorithm along with its decision. A human can check this explanation to ensure that it is an 'acceptable' explanation. We can imagine an explanation for a rejected application being "the applicant's address is in a neighborhood inhabited primarily by people of dark complexion and the applicant is short". This is a terrible explanation that does not justify the rejection of a loan application. The algorithm was clearly trained on data that had high correlation rates between loan rejection and short people with dark complexions. These are clearly irrelevant considerations and the result of the ML algorithm should be rejected.

On the other hand, we can imagine a decision and explanation by the ML for an accepted loan application that is 'acceptable'. The explanation might be something like "the applicant has a low debt to savings ratio and a high income to rent ratio". These both seem to be relevant considerations when deciding whether to accept a loan application. The problem with all of this is that in order for the explanation given by explainable AI to be useful we must have a human capable of knowing which considerations are acceptable and which are not. If we already know which considerations are acceptable, then there is no reason to use ML in the first place. We could simply hard-code the considerations into an algorithm—giving us an automated decision using pre-approved, transparent, reasoning. In this example, we would definitely not include considerations like height and race. We would instead have considerations like debt to saving ration and income to rent ratio.

For an explanation of a decision made by an ML algorithm to be useful we already need to know what counts as an acceptable consideration for that decision. For example, we can imagine an ML algorithm that could make a modern painting and could give us an explanation for each brushstroke. Since there is no agreed-upon list of considerations that 'justify' a brush stroke in the context of modern painting, it would be a useless explanation. We could do nothing with that explanation with regard to the decision it made (e.g. reject its decision). Here, the reader may think that the explanation would still be useful. We may just be curious to know why the algorithm did what it did. Furthermore, if one was a modernist painter, then this information could be used to help them become a better painter.

There is no doubt some truth to this. Explainable AI could be used to find correlations that should serve as considerations regarding the class of decisions at hand. However, explicability in these scenarios is very different. Now the explanations proffered by explainable AI are not justifying explanations—they cannot be used to justify a specific decision. For example, if a medical diagnosis algorithm used as a consideration that the patient's eyes were a very specific color, we would not immediately be able to tell if this was an acceptable reason or not. This may cause us to test the hypothesis that this specific eye color was strongly correlated with the

[17] This is unlikely as there is widespread acknowledgement that explainability and power conflict and must be traded off in the context of AI.

diagnosis. If this eye color is indeed indicative (to a medically significant level) then the algorithm's explanation would have contributed to the scientific and medical community by coming up with a consideration we had not thought of before. This consideration can now be used by GOFAI and/or doctors to make future diagnoses. In cases like these, we would no longer be checking an algorithm's explanation to ensure that it conforms to our view of what's acceptable; rather, the explanation would hopefully point us towards acceptable considerations we hadn't thought of before. Once we have these new considerations, then we could just hard code them into traditional automation algorithms rather than let the ML algorithm take the role of decision-maker.

6 Conclusion

If my arguments in this paper are on the right track, then we will find the solution for the opacity of ML by using ML for roles, decisions or actions which do not have the property of 'requiring an explanation'. This solution may seem, at first glance, to restrict ML to playing games. If games are the only things without the property of 'requiring an explanation' that ML can do well then this would be true. However, ML has had much success to date in contexts-like healthcare—that have ethical and societal import. Much of this success has been making decisions that do not require explanations. Detecting cancerous moles is one such example. An algorithm can take a picture of a mole and classify it as malignant or not. The consequences of this decision are simply a biopsy if the mole is labeled as malignant. This algorithm also outperforms dermatologists at such classification (Esteva et al. 2017; Presse, 2018). The initial classification by a doctor is done by simply looking at the mole—and although there are certain 'rules of thumb' regarding size, color, and shape, it is difficult to articulate what malignant moles look like. A doctor is not required to explain their decision. An algorithm should not be required to either—especially when it outperforms human beings at the task.

One difficulty that arises with algorithms that perform tasks like the one above is that they may still be biased and indirectly harm a group of people. Although it seems that the algorithm has a net benefit to society in that it outperforms doctors at labeling moles malignant—this benefit may not be the same for all groups of people. In this case, the algorithm performs poorly on those with a dark complexion (Lashbrook 2018). Note that this does not have anything to do with explicability as used in principles for AI. The algorithm is not using skin color as a consideration for determining whether a mole is malignant; rather, the algorithm is not very good at labeling moles on patients with dark complexions. To take a simpler example, when an individual practices a presentation before a conference they may be able to pace the presentation well, speak clearly, and not lose their place. When it comes to the actual presentation in front of a group of people, they could still perform much worse. They speak too fast—causing them to end too early—and lose their place which causes them to skip over slides because they cannot remember what they were supposed to say for them. They did not decide to perform badly because they were in front of a group of people. Quite the contrary—they made a conscious effort

to perform their best. They are just not very good at presenting in front of people. The source of their problem—and the problem with many ML algorithms—is not in the explanation of the decision but in the efficacy of its decisions/actions given different contexts and inputs.

In this article, I have argued that the property of 'requiring an explanation' belongs to the decisions and actions themselves—not the entity performing the action or decision. When we direct our attention to those decisions and actions, we can better decide in which contexts and roles we should be using ML algorithms. Furthermore, in showing that there is a catch 22 for explicable ML algorithms, it is argued that the reason for making explicable AI is an epistemic one—not a moral obligation. The only way to use explicable ML to solve the moral issue of algorithmic opacity is if we have already figured out the acceptable considerations for making the decision or performing the action at hand. If we already have those acceptable considerations, there is no need to use ML in the first place.

Acknowledgements I would like to thank Aimee van Wynsberghe, Seumas Miller, Mark Alfano, Jonas Feltes, Madelaine Ley, Tom Coggins, and two anonymous reviewers for their helpful feedback on earlier drafts.

Funding The research benefited from the activities undertaken in the European Research Council advanced grant project "Collective Responsibility and Counterterrorism" awarded to Professor Seumas Miller.

References

Ahmed, M. (2018). Aided by Palantir, the LAPD uses predictive policing to monitor specific people and neighborhoods. *The Intercept*. Retrieved October 11, 2018, from The Intercept website: https://theintercept.com/2018/05/11/predictive-policing-surveillance-los-angeles/.

AI at Google: Our principles. (2018). *Google*. Retrieved January 14, 2019, from Google website: https://www.blog.google/technology/ai/ai-principles/.

AI Principles. (2017). *Future of Life Institute*. Retrieved January 14, 2019, from Future of Life Institute website: https://futureoflife.org/ai-principles/.

AI Universal Guidelines—thepublicvoice.org. (2018). *The Public Voice*. Retrieved January 14, 2019, from https://thepublicvoice.org/ai-universal-guidelines/.

Article 36. (2015). Killing by machine: Key issues for understanding meaningful human control website. *Article 36*. Retrieved April 4, 2019, from Article 36 website: http://www.article36.org/autonomous-weapons/killing-by-machine-key-issues-for-understanding-meaningful-human-control/.

Barry-Jester, A., Casselman, B., & Goldstein, D. (2015). The new science of sentencing. *The Marshall Project*. Retrieved January 17, 2019, from The Marshall Project website: https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing.

Berk, R. A., Sorenson, S. B., & Barnes, G. (2016). Forecasting domestic violence: A machine learning approach to help inform arraignment decisions. *Journal of Empirical Legal Studies, 13*(1), 94–115. https://doi.org/10.1111/jels.12098.

Bridle, J. (2019). *New dark age: Technology and the end of the future* (Reprint edition). Verso.

Dancy, J. (2004). *Practical reality*. Oxford: Oxford University Press.

Darwall, S. (2003). Desires, reasons, and causes. *Philosophy and Phenomenological Research, 67*(2), 436–443. https://doi.org/10.1111/j.1933-1592.2003.tb00300.x.

de Bruijne, M. (2016). Machine learning approaches in medical image analysis: From detection to diagnosis. *Medical Image Analysis, 33,* 94–97. https://doi.org/10.1016/j.media.2016.06.032.

Dhar, J., & Ranganathan, A. (2015). Machine learning capabilities in medical diagnosis applications: Computational results for hepatitis disease. *International Journal of Biomedical Engineering and Technology, 17*(4), 330–340. https://doi.org/10.1504/IJBET.2015.069398.

Edwards, L., & Veale, M. (2017). Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. *Duke Law & Technology Review, 16,* 18.

Ekelhof, M. (2019). Moving beyond semantics on autonomous weapons: Meaningful human control in operation. *Global Policy, 10,* 343–348. https://doi.org/10.1111/1758-5899.12665.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2017). Runaway feedback loops in predictive policing. In *Proceedings of machine learning research*, 81, 1–12. Retrieved from http://arxiv.org/abs/1706.09847.

Erickson, B. J., Korfiatis, P., Akkus, Z., & Kline, T. L. (2017). Machine learning for medical imaging. *RadioGraphics, 37*(2), 505–515. https://doi.org/10.1148/rg.2017160130.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115–118. https://doi.org/10.1038/nature21056.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … Vayena, E. (2018). AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5.

France-Presse, A. (2017). World's best Go player flummoxed by Google's 'godlike' AlphaGo AI. *The Guardian*. Retrieved May 22, 2019, from https://www.theguardian.com/technology/2017/may/23/alphago-google-ai-beats-ke-jie-china-go.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *IEEE 5th international conference on data science and advanced analytics (DSAA)*, 80–89. https://doi.org/10.1109/DSAA.2018.00018.

Heikoop, D. D., Hagenzieker, M., Mecacci, G., Calvert, S., Sio, F. S. D., & van Arem, B. (2019). Human behaviour with automated driving systems: A quantitative framework for meaningful human control. *Theoretical Issues in Ergonomics Science*. https://doi.org/10.1080/1463922X.2019.1574931.

Heller, J. (2011). *Catch-22*. New York: Random House.

Horowitz, M. C., & Scharre, P. (2015). Meaningful human control in weapons systems: A primer. *Center for a New American Security*. Retrieved September 2, 2019, from Center for a New American Security website: https://s3.amazonaws.com/files.cnas.org/documents/Ethical_Autonomy_Working_Paper_031315.pdf?mtime=20160906082316.

Joh, E. E. (2017). Feeding the machine: Policing, crime data, & algorithms. *William & Mary Bill of Rights Journal, 26,* 287.

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology, 8*(4), 195–204. https://doi.org/10.1007/s10676-006-9111-5.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415.

Kasperkevic, J. (2015). Google says sorry for racist auto-tag in photo app. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2015/jul/01/google-sorry-racist-auto-tag-photo-app.

Kuang, C. (2017). Can A.I. be taught to explain itself? *The New York Times*. Retrieved from https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html.

Lashbrook, A. (2018). AI-driven dermatology could leave dark-skinned patients behind. *The Atlantic*. Retrieved October 3, 2018, from The Atlantic website: https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/.

Metz, C. (2016). In two moves, AlphaGo and Lee Sedol redefined the future. *Wired*. Retrieved from https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/.

Morrell, A. (2018). Citigroup has inked a deal with an AI-powered fintech to help flag suspicious payments and safeguard a $4 trillion daily operation. *Business Insider*. Retrieved January 17, 2019, from Business Insider website: https://www.businessinsider.com/citi-has-inked-a-deal-with-an-ai-powered-fintech-feedzai-2018-12.

Nadella, S. (2016). Microsoft's CEO explores how humans and A.I. Can solve society's challenges—together. *Slate*. Retrieved January 14, 2019, from Slate Magazine website: https://slate.com/techn

ology/2016/06/microsoft-ceo-satya-nadella-humans-and-a-i-can-work-together-to-solve-societys-challenges.html.
Partnership on AI. (2019). About page. Retrieved January 16, 2019, from The Partnership on AI website: https://www.partnershiponai.org/about/.
Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Cambridge, MA: Harvard University Press.
Presse, A. F. (2018). Computer learns to detect skin cancer more accurately than doctors. *The Guardian*. Retrieved from https://www.theguardian.com/society/2018/may/29/skin-cancer-computer-learns-to-detect-skin-cancer-more-accurately-than-a-doctor.
Robbins, S. (2019). AI and the path to envelopment: Knowledge as a first step towards the responsible regulation and use of AI-powered machines. *AI & SOCIETY*. https://doi.org/10.1007/s00146-019-00891-1.
Robbins, S., & Henschke, A. (2017). The value of transparency: Bulk data and authoritarianism. *Surveillance & Society., 15*(3/4), 582–589. https://doi.org/10.24908/ss.v15i3/4.6606.
Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*. https://doi.org/10.3389/frobt.2018.00015.
Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., … Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529*(7587), 484–489. https://doi.org/10.1038/nature16961.
UNI Global Union. (2018). 10 principles for ethical AI. Retrieved April 10, 2019, from UNI Global Union website: http://www.thefutureworldofwork.org/opinions/10-principles-for-ethical-ai/.
United States Department of Defense. (2012). *Department of defense directive on autonomous weapons systems*. Retrieved September 2, 2019, from https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.pdf.
Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. *Proceedings of the Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3287560.3287566.
Vincent, J. (2018). AI that detects cardiac arrests during emergency calls will be tested across Europe this summer. *The Verge*. Retrieved May 23, 2018, from The Verge website: https://www.theverge.com/2018/4/25/17278994/ai-cardiac-arrest-corti-emergency-call-response.
Wachter, S., Mittelstadt, B., & Floridi, L. (2016). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law, 7*(2), 76–99. https://doi.org/10.1093/idpl/ipx005.
Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2). Retrieved January 16, 2019, from http://arxiv.org/abs/1711.00399.
Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., … Schwartz, O. (2018). *AI Now*. Retrieved January 16, 2019, from AI Now Institute website: https://ainowinstitute.org/AI_Now_2018_Report.html.